J. Clin. Chem. Clin. Biochem.
Vol. 17, 1979, pp. 565–572

Falsch positive Werte bei der Vielfachanalyse: Die Abschätzung ihrer Häufigkeit mit der *Sylvester*schen Formel und ihre Reduktion durch eine multivariate Testgröße

Von *W. Rehpenning, K. Harm, Astrid Domesle* und *K. D. Voigt*

Abteilung für Mathematik in der Medizin (Prof. Dr. J. Berger) und Zentrallabor der Medizinischen Kliniken (Prof. Dr. K. D. Voigt) des Universitäts-Krankenhauses Hamburg-Eppendorf

(Eingegangen am 1. Dezember 1978/2. Mai 1979)

Zusammenfassung: Die Anwendung der Binomialverteilung zur Bestimmung der Anzahl „falsch positiver Werte" bei der Profilanalytik ist nicht zulässig wegen der teilweise hohen Korrelationen zwischen den klinisch-chemischen Kenngrößen. Durch Verallgemeinerung haben wir deshalb versucht, mit Hilfe der *Sylvester*schen Formel genauere Abschätzungen unter Berücksichtigung der Korrelationen zu erhalten. Wir benutzen multivariate Testgrößen zur Beurteilung des Datenvektors eines Patienten. Wir kommen zu dem Schluß, daß eine Neudefinition des Begriffes des „falsch positiven Wertes" notwendig ist, wenn mehrere Laborkenngrößen gleichzeitig betrachtet werden. Die multivariate Technik führt einerseits zu einer erheblichen Reduktion der zu Unrecht als „falsch positiv" interpretierten Werte und andererseits erlaubt sie auch die Entdeckung versteckter unplausibler Wertekonstellationen.

Falsely positive values in multi-channel analysis: An estimation of their frequency with Sylvester's formula and their reduction by a multivariate test quantity

Summary: The application of the binomial distribution for determining the number of "falsely positive values" is not suitable because of the partly strong correlations between the clinical chemical parameters. We have therefore tried to obtain more precise estimations for the number of falsely positive values by using a generalisation from *Sylvester*'s formula and taking into account the correlations. We use multivariate statistics for testing a patient's data vector. We come to the conclusion that it is necessary to introduce a new definition of the term "falsely positive value" when several laboratory parameters are simultaneously considered. The multivariate technique leads firstly to a significant reduction of the erroneously interpreted "falsely positive values" and secondly allows the detection of hidden implausible constellations of values.

Einleitung

In einer kürzlich durchgeführten Studie haben wir die Anzahl der falsch positiven Werte bei der Vielfachanalyse untersucht und dabei Abweichungen von den theoretisch nach der Binomialverteilung zu erwartenden Häufigkeiten gefunden (1). Hierbei hatten wir einen Wertesatz in einer vorläufigen Definition als falsch positiv angesehen, wenn mindestens eine Kenngröße außerhalb ihres entsprechenden 95%-Referenzbereiches liegt. Wenn man annimmt, daß die Kenngrößen voneinander unabhängig sind, läßt sich die erwartete Anzahl von falsch positiven Werten mit Hilfe der Binomialverteilung berechnen. Wir konnten nachweisen, daß die Anwendung der Binomialformel zur Bestimmung des Prozentsatzes falsch positiver Werte einerseits zu einer Überschätzung der Gesamtzahl von Profilen mit falsch positiven Werten führt, andererseits aber zu einer Unterschätzung der Anzahl von Profilen mit Werten, die gleichzeitig außerhalb der Referenzbereiche liegen. Die Erklärung hierfür ist darin zu suchen, daß die Binomialformel bei erheblichen Korrelationen zwischen den Kenngrößen nicht geeignet ist, worauf schon *Büttner* (2) hingewiesen hat.

Wir haben daher anhand von mathematischen Modellrechnungen versucht, die Problematik der falsch positiven Werte unter Berücksichtigung der Korrelationen zwischen den Kenngrößen weiter zu bearbeiten. Hierbei konnten wir eine von *Sylvester* stammende Formel mit Vorteil benutzen (3). Es wurde klar, daß sowohl bei der Qualitäts- und Plausibilitätskontrolle als auch bei der allgemeinen Beurteilung von falsch positiven und extremen Datensätzen der Einsatz multivariater Techniken unerläßlich ist. Dieses gilt insbesondere für eine wirklich sinnvolle Definition von Referenzbereichen, die sich ganz wesentlich von den üblicherweise benutzten univariaten Referenzbereichen unterscheiden. Auf diese

0340-076X/79/0017-0565$02.00

Notwendigkeit wurde bereits von verschiedenen Autoren hingewiesen (4, 5). Auch diese Problematik soll im folgenden berücksichtigt werden.

Material und Methoden

Für die Auswertung standen uns die Daten von 313 Referenzpersonen zur Verfügung, die im Jahre 1976 im Rahmen der routinemäßigen Personaluntersuchung im Universitäts-Krankenhaus Eppendorf erhoben wurden. Sämtliche Profile wurden mit einem Technicon Autoanalyzer SMA 12/60 gemessen. Jedes Profil bestand aus folgenden Kenngrößen (1): Natrium, Kalium, Chlorid, Gesamt-Eiweiß, Albumin, anorganischer Phosphor, Cholesterin, Harnstoff-Stickstoff, Calcium, Kreatinin, Bilirubin, Harnsäure.

Die Meßwerte wurden on-line vom Telefunken-Rechner TR 86 des Systems ELIAS (6) erfaßt und nach Wandlung in eine für den Großrechner TR 440 des Rechenzentrums der Universität Hamburg lesbare Form auf diesem mit Hilfe von FORTRAN-Programmen ausgewertet.

Für die Modellrechnungen erweist es sich als zweckmäßig, für die einzelnen Variablen Normalverteilungen anzunehmen. Wir setzen also voraus, daß der Datenvektor einer multivariaten Normalverteilung angehört, die durch den Mittelwertsvektor $\bar{x}$ und die Kovarianzmatrix V bestimmt ist. Wir definieren einen „falsch positiven Wertesatz" in bezug auf das Referenzkollektiv vorläufig einmal durch die Tatsache, daß mindestens ein Kennwert außerhalb seines entsprechenden Referenzbereiches liegt, der durch die Einschlußwahrscheinlichkeit $1 - \alpha$ (z. B. 0,95) gekennzeichnet ist. Bei dieser Betrachtungsweise errechnet sich die Wahrscheinlichkeit für das Auftreten eines „falsch positiven Wertesatzes" zu $P^+ = 1 - P$, wobei P gegeben ist durch das n-dimensionale Gebietsintegral bei n Variablen: $P = \int_G \varphi \, db$. Hierbei ist:

$$\varphi(x) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp\left(-\frac{1}{2}(x - \bar{x})' V^{-1}(x - \bar{x})\right) \quad \text{(Gl. 1)}$$

eine n-dimensionale Normalverteilung, x ist der Variablenvektor

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

und V die Kovarianzmatrix.

Das Gebiet G ist der n-dimensionale Würfel, der gegeben ist durch:

$$-z_\gamma \leqslant \frac{x_i - \bar{x}_i}{s_i} \leqslant z_\gamma \quad (i = 1, \ldots, n), (\gamma = 1 - \alpha/2) \quad \text{(Gl. 2)}$$

Dabei ist z_γ das entsprechende Fraktil der Normalverteilung (7). Dieses Integral ist allerdings eine Annäherung an die wahren Verhältnisse, weil wir für die unbekannten Korrelationen die empirischen Korrelationen benutzt haben. Für den Fall, daß die Variablen voneinander unabhängig sind, zerfällt das Integral in ein Produkt von n einfachen Integralen, so daß sich ergibt:

$$P = \left(\int_{-z_\gamma}^{z_\gamma} \varphi(x)\, dx\right)^n = (1 - \alpha)^n, \quad \text{(Gl. 3)}$$

d. h. eine Reduktion auf die Binomialformel. Da aber die Korrelationen nicht vernachlässigt werden können (1), wird an dieser Stelle ersichtlich, daß die Binomialformel nicht anwendbar ist.

Das Integral (Gl. 1) stellt also die Wahrscheinlichkeit dafür dar, daß alle Kenngrößen in ihrem Referenzbereich liegen, d. h., daß der Wertesatz nicht falsch positiv ist. Weil das Integral weder formelmäßig darstellbar noch mit einem vernünftigen Aufwand numerisch auswertbar ist, in unserem Fall wäre bei dem SMA 12/60-Profil n = 12, mußte nach anderen Möglichkeiten gesucht werden. Hier kommt uns eine Formel aus der Wahrscheinlichkeitsrechnung zur Hilfe, die nach *Sylvester* benannt ist und die folgendes aussagt:

Nehmen wir an, wir hätten Ereignisse $A_1, A_2, \ldots, A_n$ mit den Wahrscheinlichkeiten $P(A_1), P(A_2), \ldots, P(A_n)$. Wenn wir dann die Wahrscheinlichkeit dafür wissen wollen, daß mindestens eines dieser Ereignisse eintritt, so gilt die Formel von *Sylvester*:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j}^{n} P(A_i A_j) + \sum_{i<j<k}^{n} P(A_i A_j A_k) - + \ldots + (-1)^{n-1} P(A_1 A_2 \ldots A_n). \quad \text{(Gl. 4)}$$

Hierbei bedeutet z. B. $P(A_i A_j)$ die Wahrscheinlichkeit dafür, daß die Ereignisse A_i und A_j gleichzeitig eintreten. Entsprechendes gilt für die anderen Summen. Wichtig ist, daß die rechts stehende Summe die gesuchte Wahrscheinlichkeit alternierend einschachtelt, wenn man sie sukzessive nach den einzelnen Summen berechnet. Diese Beziehung gilt ganz allgemein.

Diese Gleichung läßt sich nun mit einer leichten Modifikation auf unser Problem anwenden. Bezeichnen wir mit A_i das Ereignis „Der Wert der i-ten Kenngröße liegt im Referenzbereich" (i = 1, ..., n), dann läßt sich die Tatsache, daß ein falsch positiver Wertesatz vorliegt, durch das Ereignis ausdrücken:

$$B = A_1^c \cup A_2^c \cup \ldots \cup A_n^c. \quad \text{(Gl. 5)}$$

Hierbei bedeutet A_i^c: „Die i-te Kenngröße liegt *nicht* im Referenzbereich".

Jetzt läßt sich durch die mehrmalige Anwendung der *Sylvester*schen Formel, wie in Abbildung 1 angedeutet, die gesuchte Wahrscheinlichkeit P(B) für das Ereignis B auf die Wahrscheinlichkeiten zurückführen, daß jeweils ein, zwei oder mehr Kanäle im Referenzbereich liegen und zwar losgelöst von dem Verhalten der anderen. Diese Wahrscheinlichkeiten lassen sich durch numerische Integration berechnen, wobei man aus praktischen Gründen die *Sylvester*sche Reihe abbrechen muß. Da aber die Summenglieder immer mehr abnehmen und den wahren Wert alternierend einschließen (8), gelangt man zu Abschätzungen für den wahren Integralwert. Bei dieser Spezialisierung stellen die Ausdrücke $A_i A_j A_k$ usw. die Ereignisse dar, daß der i-te, j-te und k-te Kanal gleichzeitig im zugehörigen Referenzbereich liegen. Die Korrelationen zwischen den einzelnen Kenngrößen, für die wir empirische *Pearson*sche Korrelationskoeffizienten benutzt haben, gehen bei der Berechnung der Integrale, z. B. $P(A_i A_j A_k)$ natürlich wesentlich ein.

$$P\left(\bigcup_{i=1}^{N} A_i^c\right) = \sum_{i=1}^{N} P(A_i^c) - \sum_{i<j}^{N} P(A_i^c A_j^c) + \sum_{i<j<k}^{N} P(A_i^c A_j^c A_k^c) - + \cdots$$

$$P(A_i^c) = 1 - P(A_i)$$

$$P(A_i^c A_j^c) = 1 - P(A_i) - P(A_j) + P(A_i A_j)$$

$$P(A_i^c A_j^c A_k^c) = 1 - P(A_i) - P(A_j) - P(A_k) + P(A_i A_j) + P(A_i A_k) + P(A_j A_k) - P(A_i A_j A_k)$$

Abb. 1. Die *Sylvester*sche Formel in der von uns benutzten Form. Sie gibt die Wahrscheinlichkeit für eine Kombination von Ereignissen an. Einzelheiten siehe Text.

Tab. 1. Integrale über die zweidimensionale Normalverteilung.

r	J	
	z = 1,9600	z = 2,5758
0,00	0,902508	0,980098
0,05	0,902574	0,980105
0,10	0,902771	0,980126
0,15	0,903099	0,980162
0,20	0,903560	0,980214
0,25	0,904155	0,980284
0,30	0,904886	0,980373
0,35	0,905756	0,980485
0,40	0,906770	0,980622
0,45	0,907935	0,980790
0,50	0,909261	0,980991
0,55	0,910762	0,981233
0,60	0,912458	0,981521
0,65	0,914376	0,981865
0,70	0,916555	0,982277
0,75	0,919051	0,982771
0,80	0,921948	0,983372
0,85	0,925390	0,984118
0,90	0,929652	0,985079
0,95	0,935435	0,986436

Um die Methode allgemein anwendbar zu machen und um jeden mit der Durchführung von Mehrkanalanalysen betrauten Klinischen Chemiker oder Laboratoriumsmediziner in den Stand zu versetzen, die Zahl der falsch positiven Datensätze bei seiner speziellen Kenngrößen- und Datenkonstellation größenordnungsmäßig richtig abzuschätzen, haben wir die mehrdimensionalen Integrale bis zur dritten Dimension für verschiedene Korrelationskoeffizienten berechnet, die in den Tabellen 1 bis 3 dargestellt sind. Dabei muß der Benutzer in einem ersten Schritt die Korrelationsmatrix seiner Parameter berechnen und dann die Abschätzungen für seine Wahrscheinlichkeiten mit Hilfe der *Sylvester*schen Formel vornehmen, wobei die Integrale nach eventueller Interpolation den Tabellen entnommen werden können.

Die Integrale über die zweidimensionale Normalverteilung sind in Tabelle 1 dargestellt. Die Tabellen 2 und 3 enthalten die Werte der Integrale über die trivariate Normalverteilung über den Würfel $-1{,}9600 \leqslant z_i \leqslant 1{,}9600$ bzw. $-2{,}5758 \leqslant z_i \leqslant 2{,}5758$, $(i = 1, 2, 3)$. Die Werte sind auf sechs Stellen genau. Aus praktischen Gründen haben wir uns auf Korrelationen mit $|r| \leqslant 0{,}6$ beschränkt. Bei der Benutzung der Tafeln ist zu beachten, daß aus Symmetriegründen gewisse Vertauschungsrelationen bestehen. So gilt bei offensichtlicher Bezeichnungsweise für die Integrale:

$$J(a, b, c) = J(a, c, b) = J(b, a, c) = J(b, c, a) = J(c, a, b) = J(c, b, a). \quad \text{(Gl. 6)}$$

Zusätzlich gilt:

$$J(-a, -b, c) = J(a, b, c) \text{ und } J(0, b, -c) = J(0, b, c). \quad \text{(Gl. 7)}$$

Außerdem sei angemerkt, daß gewisse Konstellationen von Korrelationskoeffizienten nicht möglich sind, z. B. gibt es keine Korrelationsmatrix mit den Werten $r_{1,2} = 0{,}4$, $r_{1,3} = 0{,}5$ und $r_{2,3} = -0{,}6$. Das liegt daran, daß die entsprechende Determinante $|V|$ aus Gleichung 1 negativ ist. Die Konstellation $r_{1,2} = 0{,}5$, $r_{1,3} = 0{,}5$ und $r_{2,3} = -0{,}5$ ist singulär. Für den Anwendungsfall muß zwischen den Tafeln interpoliert werden.

Für die prinzipiell schwierige dreidimensionale Interpolation schlagen wir folgendes einfache Verfahren vor:

Die Integralwerte sind auf einem würfelförmigen Raster gegeben. Die Inkremente in Richtung der drei Achsen innerhalb eines bestimmten Würfels in Einheiten der Kantenlänge seien: h, k, l.

Die tabellierten Integralwerte in den Ecken des Würfels seien in sinnvoller Reihenfolge mit J_{000}, J_{100}, J_{010}, J_{001}, J_{110}, J_{101}, J_{011} und J_{111} bezeichnet. Außerdem sei $h' = 1 - h$, $k' = 1 - k$ und $l' = 1 - l$. Dann ergibt sich für den gesuchten Integralwert:

$$J = h'k'l'J_{000} + hk'l'J_{100} + h'kl'J_{010} + h'k'lJ_{001} + hkl'J_{110} + hk'lJ_{101} + h'klJ_{011} + hklJ_{111}. \quad \text{(Gl. 8)}$$

Ein Beispiel mag die Interpolationsmethode verdeutlichen: Es sei nach bereits erfolgter Vertauschung $r_{1,2} = 0{,}22$, $r_{1,3} = 0{,}37$, $r_{2,3} = -0{,}54$. Dann hat die linke untere Ecke des Würfels die Koordinaten $\begin{pmatrix} 0{,}2 \\ 0{,}3 \\ -0{,}6 \end{pmatrix}$. Weil die Kantenlänge des Würfels gleich 0,1 ist, ergibt sich: $h = 0{,}2$, $k = 0{,}7$, $l = 0{,}6$, $h' = 0{,}8$, $k' = 0{,}3$, $l' = 0{,}4$. Aus der Tabelle entnimmt man z. B. für den 95%-Bereich die Integralwerte:

$J_{000} = 0{,}870513$ $\quad$ $J_{110} = 0{,}873922$
$J_{100} = 0{,}871964$ $\quad$ $J_{101} = 0{,}868797$
$J_{010} = 0{,}872427$ $\quad$ $J_{011} = 0{,}869271$
$J_{001} = 0{,}867382$ $\quad$ $J_{111} = 0{,}870731.$

Damit ergibt sich $J = 0{,}870256$. Der wahre Wert ist $J = 0{,}870067$.

Weitere Beispiele zeigen, daß die Genauigkeit der nach dieser Formel interpolierten Werte etwa 1 bis 2 Einheiten der vierten Dezimale ausmacht, eine für diese Zwecke ausreichende Genauigkeit.

Zur Erläuterung der Methodik bringen wir noch ein einfaches Beispiel, bei dem keine Interpolation erforderlich ist:

Es sei $n = 4$, die Korrelationsmatrix sei gegeben durch $r_{1,2} = 0{,}6$, $r_{1,3} = 0{,}4$, $r_{1,4} = 0{,}3$, $r_{2,3} = -0{,}3$, $r_{2,4} = 0{,}4$, $r_{3,4} = -0{,}4$. Wir stellen uns die Aufgabe, anhand der Integraltafeln die erwartete Häufigkeit falsch positiver Werte für $1 - \alpha = 0{,}95$ zu bestimmen. Tabelle 1 und 2 entnehmen wir die Werte:

$J_{1,2} = 0{,}912458$
$J_{1,3} = 0{,}906770$
$J_{1,4} = 0{,}904886$ $\quad$ $J_{1,2,3} = 0{,}873922$
$J_{2,3} = 0{,}904886$ $\quad$ $J_{1,2,4} = 0{,}871985$
$J_{2,4} = 0{,}906770$ $\quad$ $J_{1,3,4} = 0{,}868238$
$J_{3,4} = 0{,}906770$ $\quad$ $J_{2,3,4} = 0{,}867014.$

Bezeichnen wir die bei der *Sylvester*schen Formel (Abb. 1) benötigten Summen der Dimension nach geordnet mit S_1, S_2 und S_3, dann ergibt sich der Reihe nach:

$$S_1 = \sum_{i=1}^{4} P(A_i^c) = \sum_{i=1}^{4} 1 - P(A_i) = 4 - 4 \cdot 0{,}95 = 0{,}200000$$

$$S_2 = \sum_{i<j}^{4} P(A_i^c A_j^c) = \sum_{i<j}^{4} 1 - P(A_i) - P(A_j) + P(A_i A_j) = 6(1 - 2 \cdot 0{,}95) + J_{1,2} + J_{1,3} + J_{1,4} + J_{2,3} + J_{2,4} + J_{3,4} = 0{,}042540$$

$$S_3 = \sum_{i<j<k}^{4} P(A_i^c A_j^c A_k^c) = \sum_{i<j<k}^{4} 1 - P(A_i) - P(A_j) - P(A_k) + P(A_i A_j) + P(A_i A_k) + P(A_j A_k) - P(A_i A_j A_k) = 4(1 - 3 \cdot 0{,}95) + 2 \cdot 5{,}442540 - 3{,}481159 = 0{,}003921.$$

Tab. 2. Integrale über die dreidimensionale Normalverteilung (z = 1,9600)

$r_{2,3}$	J	$r_{2,3}$	J	$r_{2,3}$	J	$r_{2,3}$	J	$r_{2,3}$	J	$r_{2,3}$	J
$r_{1,2} = 0{,}0$		$r_{1,3} = 0{,}5$		$r_{1,3} = 0{,}3$		$r_{1,2} = 0{,}2$		$r_{1,3} = 0{,}5$		0,4	0,867014	−0,1	0,868293
										0,5	0,869172	0,5	0,872646
$r_{1,3} = 0{,}0$		0,5	0,870141	−0,6	0,869546	$r_{1,3} = 0{,}2$		−0,6	0,874926	0,6	0,871985	0,6	0,875306
		0,6	0,873186	−0,5	0,866460			−0,5	0,871747				
0,0	0,857386			−0,4	0,864061	−0,6	0,869122	−0,4	0,869271	$r_{1,3} = 0{,}5$		$r_{1,3} = 0{,}6$	
0,1	0,857636	$r_{1,3} = 0{,}6$		−0,3	0,862244	−0,5	0,866019	−0,3	0,867382				
0,2	0,858386			−0,2	0,860958	−0,4	0,863605	−0,2	0,866019	−0,6	0,876445	−0,4	0,875898
0,3	0,859646	0,6	0,876256	−0,1	0,860177	−0,3	0,861773	−0,1	0,865157	−0,5	0,873242	−0,3	0,873922
0,4	0,861436			0,3	0,862021	−0,2	0,860472	0,5	0,870414	−0,4	0,870731	−0,2	0,872427
0,5	0,863802	$r_{1,2} = 0{,}1$		0,4	0,863759	−0,1	0,859678	0,6	0,873272	−0,3	0,868797	−0,1	0,871395
0,6	0,866839			0,5	0,866070	0,2	0,860277			−0,2	0,867382	0,6	0,877893
		$r_{1,3} = 0{,}1$		0,6	0,869051	0,3	0,861477	$r_{1,3} = 0{,}6$		−0,1	0,866460		
$r_{1,3} = 0{,}1$						0,4	0,863203			0,5	0,871276	$r_{1,2} = 0{,}5$	
		−0,6	0,867410	$r_{1,3} = 0{,}4$		0,5	0,865501	−0,6	0,878126	0,6	0,874036		
0,1	0,857885	−0,5	0,864356			0,6	0,868463	−0,5	0,874926			$r_{1,3} = 0{,}5$	
0,2	0,858635	−0,4	0,861977	−0,6	0,871395			−0,4	0,872427	$r_{1,3} = 0{,}6$			
0,3	0,859893	−0,3	0,860177	−0,5	0,868293	$r_{1,3} = 0{,}3$		−0,3	0,870513			−0,4	0,875218
0,4	0,861682	−0,2	0,858907	−0,4	0,865885			−0,2	0,869122	−0,5	0,876445	−0,3	0,873242
0,5	0,864047	−0,1	0,858146	−0,3	0,864061	−0,6	0,870513	−0,1	0,868227	−0,4	0,873922	−0,2	0,871747
0,6	0,867083	0,1	0,858122	−0,2	0,862767	−0,5	0,867382	0,6	0,876110	−0,3	0,871964	−0,1	0,870713
		0,2	0,858859	−0,1	0,861977	−0,4	0,864945			−0,2	0,870513	0,5	0,874567
$r_{1,3} = 0{,}2$		0,3	0,860104	0,4	0,865476	−0,3	0,863094	$r_{1,2} = 0{,}3$		−0,1	0,869546	0,6	0,877127
		0,4	0,861877	0,5	0,867767	−0,2	0,861773			0,6	0,876749		
0,2	0,859382	0,5	0,864227	0,6	0,870731	−0,1	0,860958	$r_{1,3} = 0{,}3$				$r_{1,3} = 0{,}6$	
0,3	0,860638	0,6	0,867245			0,3	0,862647			$r_{1,2} = 0{,}4$			
0,4	0,862422			$r_{1,3} = 0{,}5$		0,4	0,864340	−0,6	0,871964			−0,3	0,876445
0,5	0,864784	$r_{1,3} = 0{,}2$				0,5	0,866602	−0,5	0,868797	$r_{1,3} = 0{,}4$		−0,2	0,874926
0,6	0,867820			−0,6	0,873836	0,6	0,869527	−0,4	0,866329			−0,1	0,873836
		−0,6	0,868227	−0,5	0,870713			−0,3	0,864447	−0,6	0,875898	0,6	0,879584
$r_{1,3} = 0{,}3$		−0,5	0,865157	−0,4	0,868293	$r_{1,3} = 0{,}4$		−0,2	0,863094	−0,5	0,872694		
		−0,4	0,862767	−0,3	0,866460			−0,1	0,862244	−0,4	0,870178	$r_{1,2} = 0{,}6$	
0,3	0,861888	−0,3	0,860958	−0,2	0,865157	−0,6	0,872427	0,3	0,863772	−0,3	0,868238		
0,4	0,863667	−0,2	0,859678	−0,1	0,864356	−0,5	0,869271	0,4	0,865418	−0,2	0,866815	$r_{1,3} = 0{,}6$	
0,5	0,866024	−0,1	0,858907	0,5	0,870040	−0,4	0,866815	0,5	0,867629	−0,1	0,865885		
0,6	0,869058	0,2	0,859581	0,6	0,872994	−0,3	0,864945	0,6	0,870498	0,4	0,868543	−0,2	0,878126
		0,3	0,860810			−0,2	0,863605			0,5	0,870631	−0,1	0,876993
$r_{1,3} = 0{,}4$		0,4	0,862567	$r_{1,3} = 0{,}6$		−0,1	0,862767	$r_{1,3} = 0{,}4$		0,6	0,873368	0,6	0,881909
		0,5	0,864898			0,4	0,865998						
0,4	0,865439	0,6	0,867897	−0,6	0,876993	0,5	0,868224	−0,6	0,873922	$r_{1,3} = 0{,}5$			
0,5	0,867791			−0,5	0,873836	0,6	0,871113	−0,5	0,870731				
0,6	0,870827			−0,4	0,871395			−0,4	0,868238	−0,5	0,875218		
				−0,3	0,869546			−0,3	0,866329	−0,4	0,872694		
				−0,2	0,868227			−0,2	0,864945	−0,3	0,870731		
				−0,1	0,867410			−0,1	0,864061	−0,2	0,869271		
				0,6	0,875952								

Tab. 3. Integrale über die dreidimensionale Normalverteilung (z = 2,5758)

$r_{2,3}$	J
$r_{1,2} = 0,0$	
$r_{1,3} = 0,0$	
0,0	0,970297
0,1	0,970324
0,2	0,970411
0,3	0,970569
0,4	0,970815
0,5	0,971180
0,6	0,971705
$r_{1,3} = 0,1$	
0,1	0,970352
0,2	0,970439
0,3	0,970596
0,4	0,970843
0,5	0,971207
0,6	0,971732
$r_{1,3} = 0,2$	
0,2	0,970525
0,3	0,970682
0,4	0,970928
0,5	0,971292
0,6	0,971816
$r_{1,3} = 0,3$	
0,3	0,970838
0,4	0,971083
0,5	0,971446
0,6	0,971969
$r_{1,3} = 0,4$	
0,4	0,971327
0,5	0,971689
0,6	0,972211

$r_{2,3}$	J
$r_{1,3} = 0,5$	
0,5	0,972050
0,6	0,972572
$r_{1,3} = 0,6$	
0,6	0,973096
$r_{1,2} = 0,1$	
$r_{1,3} = 0,1$	
-0,6	0,971763
-0,5	0,971237
-0,4	0,970872
-0,3	0,970625
-0,2	0,970467
-0,1	0,970380
0,1	0,970379
0,2	0,970466
0,3	0,970622
0,4	0,970867
0,5	0,971231
0,6	0,971753
$r_{1,3} = 0,2$	
-0,6	0,971851
-0,5	0,971325
-0,4	0,970960
-0,3	0,970713
-0,2	0,970555
-0,1	0,970467
0,2	0,970551
0,3	0,970706
0,4	0,970950
0,5	0,971312
0,6	0,971832

$r_{2,3}$	J
$r_{1,3} = 0,3$	
-0,6	0,972009
-0,5	0,971483
-0,4	0,971117
-0,3	0,970870
-0,2	0,970713
-0,1	0,970625
0,3	0,970860
0,4	0,971102
0,5	0,971461
0,6	0,971979
$r_{1,3} = 0,4$	
-0,6	0,972256
-0,5	0,971729
-0,4	0,971364
-0,3	0,971117
-0,2	0,970960
-0,1	0,970872
0,4	0,971342
0,5	0,971698
0,6	0,972214
$r_{1,3} = 0,5$	
-0,6	0,972622
-0,5	0,972095
-0,4	0,971729
-0,3	0,971483
-0,2	0,971325
-0,1	0,971237
0,5	0,972053
0,6	0,972567
$r_{1,3} = 0,6$	
-0,6	0,973151
-0,5	0,972622
-0,4	0,972256
-0,3	0,972009
-0,2	0,971851
-0,1	0,971763
0,6	0,973081

$r_{2,3}$	J
$r_{1,2} = 0,2$	
$r_{1,3} = 0,2$	
-0,6	0,971944
-0,5	0,971416
-0,4	0,971049
-0,3	0,970801
-0,2	0,970643
-0,1	0,970555
0,2	0,970635
0,3	0,970788
0,4	0,971029
0,5	0,971388
0,6	0,971904
$r_{1,3} = 0,3$	
-0,6	0,972104
-0,5	0,971576
-0,4	0,971209
-0,3	0,970960
-0,2	0,970801
-0,1	0,970713
0,3	0,970939
0,4	0,971177
0,5	0,971531
0,6	0,972044
$r_{1,3} = 0,4$	
-0,6	0,972354
-0,5	0,971825
-0,4	0,971457
-0,3	0,971209
-0,2	0,971049
-0,1	0,970960
0,4	0,971412
0,5	0,971762
0,6	0,972269

$r_{2,3}$	J
$r_{1,3} = 0,5$	
-0,6	0,972722
-0,5	0,972193
-0,4	0,971825
-0,3	0,971576
-0,2	0,971416
-0,1	0,971325
0,5	0,972108
0,6	0,972611
$r_{1,3} = 0,6$	
-0,6	0,973252
-0,5	0,972722
-0,4	0,972354
-0,3	0,972104
-0,2	0,971944
-0,1	0,971851
0,6	0,973111
$r_{1,2} = 0,3$	
$r_{1,3} = 0,3$	
-0,6	0,972267
-0,5	0,971737
-0,4	0,971369
-0,3	0,971120
-0,2	0,970960
-0,1	0,970870
0,3	0,971085
0,4	0,971319
0,5	0,971668
0,6	0,972173
$r_{1,3} = 0,4$	
-0,6	0,972517
-0,5	0,971987
-0,4	0,971619
-0,3	0,971369
-0,2	0,971209

$r_{2,3}$	J
-0,1	0,971117
0,4	0,971547
0,5	0,971890
0,6	0,972388
$r_{1,3} = 0,5$	
-0,6	0,972886
-0,5	0,972356
-0,4	0,971987
-0,3	0,971737
-0,2	0,971576
-0,1	0,971483
0,5	0,972225
0,6	0,972716
$r_{1,3} = 0,6$	
-0,5	0,972886
-0,4	0,972517
-0,3	0,972267
-0,2	0,972104
-0,1	0,972009
0,6	0,973199
$r_{1,2} = 0,4$	
$r_{1,3} = 0,4$	
-0,6	0,972768
-0,5	0,972238
-0,4	0,971869
-0,3	0,971619
-0,2	0,971457
-0,1	0,971364
0,4	0,971768
0,5	0,972102
0,6	0,972589

$r_{2,3}$	J
$r_{1,3} = 0,5$	
-0,5	0,972607
-0,4	0,972238
-0,3	0,971987
-0,2	0,971825
-0,1	0,971729
0,5	0,972425
0,6	0,972901
$r_{1,3} = 0,6$	
-0,4	0,972768
-0,3	0,972517
-0,2	0,972354
-0,1	0,972256
0,6	0,973364
$r_{1,2} = 0,5$	
$r_{1,3} = 0,5$	
-0,4	0,972607
-0,3	0,972356
-0,2	0,972193
-0,1	0,972095
0,5	0,972734
0,6	0,973194
$r_{1,3} = 0,6$	
-0,3	0,972886
-0,2	0,972722
-0,1	0,972622
0,6	0,973635
$r_{1,2} = 0,6$	
$r_{1,3} = 0,6$	
-0,2	0,973252
-0,1	0,973151
0,6	0,974053

Tab. 4. *Pearson*-Korrelationskoeffizienten zwischen den Kenngrößen des SMA 12/60-Profils im Referenzkollektiv.

	Natrium	Kalium	Chlorid	Ges. Eiweiß	Albumin	Anorg. P	Cholest.	Harnst.-N	Calcium	Kreatinin	Bilirubin	Harnsäure
Natrium	1,0000	0,1826	0,4975	0,0660	0,1241	0,1154	−0,0204	0,1243	0,2575	0,2110	0,0537	0,2253
Kalium		1,0000	0,1499	0,0904	0,0250	0,1214	0,1412	0,1159	0,1805	0,0446	−0,0599	0,0328
Chlorid			1,0000	−0,1953	−0,2001	0,0616	−0,2051	−0,0585	−0,1900	−0,0078	−0,1353	−0,0700
Gesamt-Eiweiß				1,0000	0,6534	−0,0006	−0,0566	0,0522	0,5340	0,1321	0,1407	0,0170
Albumin					1,0000	0,0811	−0,0431	0,1942	0,5971	0,1869	0,1774	0,1445
Anorg.-Phosphor						1,0000	−0,1196	−0,0108	0,1108	−0,1188	−0,0147	−0,1135
Cholesterin							1,0000	0,0863	0,1457	0,1277	−0,0895	0,2297
Harnstoff-N								1,0000	0,1182	0,3742	0,0467	0,2800
Calcium									1,0000	0,1630	0,1462	0,1444
Kreatinin										1,0000	0,1474	0,5025
Bilirubin											1,0000	0,0990
Harnsäure												1,0000

Hierbei sind die Zahlen 5,442540 bzw. 3,481159 die Summen der zwei- bzw. dreidimensionalen Integrale wie oben angegeben. Damit ergibt sich die gesuchte Wahrscheinlichkeit

$$P(B) = 0{,}200000 - 0{,}042540 + 0{,}003921 = 0{,}161381.$$

Wir haben zusätzlich den wahren Wert durch vierdimensionale Integration errechnet und erhielten 0,161329, also eine Übereinstimmung in vier Dezimalen.

Die hier beschriebenen Methoden haben wir auf unser Referenzkollektiv angewandt. Die Korrelationen zwischen den einzelnen Variablen zeigt Tabelle 4. Für die Berechnung der Wahrscheinlichkeiten falsch positiver Werte haben wir aber nicht die Integraltafeln benutzt, sondern haben alle vorkommenden Integrale unter Benutzung der Korrelationen numerisch ausgewertet. Dabei sind wir bis zur fünften Dimension gegangen. Als Referenzbereiche wählten wir den 95%- und den 99%-Bereich, haben also mit den Integrationsgrenzen $z = \pm 1{,}9600$ und $z = \pm 2{,}5758$ ($\gamma = 0{,}975$ bzw. $\gamma = 0{,}995$) gearbeitet. Da die Integrationsbereiche mehrdimensionale Rechtecke sind, lassen sich die mehrdimensionalen Integrale ohne Schwierigkeiten in mehrfache Integrale umwandeln. Für die Integrationen benutzen wir angepaßte, einfach genaue Versionen des Integrators „AIND" (9). Diese Integrationsprogramme lieferten die Integrale bis zur dritten Dimension in vertretbarer Rechenzeit bis auf neun geltende Ziffern genau. Da sowohl die Anzahl der Integrale als auch die Rechenzeit pro Integral unverhältnismäßig stark anwachsen, mußten für höherdimensionale Integrale schnellere Methoden angewandt werden, die keine Fehlerabschätzungen mehr enthalten. Hier bewährte sich eine mehrdimensionale Erweiterung der 12-Punkte-*Gauss*-Formel (10). Weil die Berechnung aller fünfdimensionalen Integrale zu zeitaufwendig gewesen wäre, haben wir uns darauf beschränkt, stichprobenartig etwa 40 Integrale auszuwerten und den Gesamtbeitrag zu extrapolieren.

Ergebnisse und Diskussion

Tabelle 5 enthält die bei der Anwendung der *Sylvester*schen Formel auftretenden Summenglieder, die wir der Dimension nach geordnet mit S_1 bis S_5 bezeichnen wollen. Die Werte in Spalte 2 wurden durch numerische Integration gewonnen. Spalte 4 enthält die entsprechenden Werte, die man bei Zugrundelegung einer Binomialverteilung erhalten würde. Wie man sieht, weichen die Werte beträchtlich voneinander ab. Für den 99%-Bereich wurde nur bis zur dritten Dimension gerechnet, weil die Integrale vernachlässigbar klein werden. Weiter fällt auf, daß die Glieder der *Sylvester*schen Reihe wesentlich langsamer abnehmen als für unkorrelierte Größen zu erwarten.

Tabelle 6 zeigt die theoretischen und beobachteten prozentualen Häufigkeiten von Profilen mit Werten außer-

Tab. 5. Summenglieder (S) bei der *Sylvester*schen Formel mit und ohne Berücksichtigung der Korrelationen zwischen den Kenngrößen bei den Referenzpersonen.

Dimension des Integrals	korreliert		unkorreliert	
	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
1	0,60000	0,12000	0,60000	0,12000
2	0,24036	0,01622	0,16500	0,00660
3	0,07697	0,00227	0,02750	0,00022
4	0,02007	–	0,00309	–
5	0,00479	–	0,00025	–

Tab. 6. Prozentuale Häufigkeit von Profilen mit Werten außerhalb der 95%-Referenzbereiche.

Anzahl der Kenngrößen außerhalb der 95%-Referenzbereiche	Prozentualer Anteil von Profilen		
	Binomialverteilung	*Sylvester*sche Formel	beobachtet
1	34,1	29,4	27,2
2	9,9	8,2	7,1
⩾ 3	2,0	4,5	5,1
Gesamt	46,0	42,1	39,4

halb der 95%-Referenzbereiche. Für die Berechnung wurde folgende Formel zugrunde gelegt (8):

$$P_m = S_m - \binom{m+1}{m} S_{m+1} + \binom{m+2}{m} S_{m+2} - \binom{m+3}{m} S_{m+3} + - \ldots \pm \binom{n}{m} S_{n'} \qquad \text{(Gl. 9)}$$

die auch wieder nach S_5 abgebrochen wurde. Hierbei bedeutet P_m die Wahrscheinlichkeit dafür, daß bei einem Profil genau m Werte außerhalb der 95%-Referenzbereiche liegen. Hierbei mußten wir uns auf $m \leqslant 2$ beschränken, weil die Summen wegen der stark wachsenden Koeffizienten stärker oszillieren als bei der einfachen *Sylvester*schen Formel. Insgesamt zeigt sich, daß die mit der *Sylvester*schen Formel unter Berücksichtigung der Korrelationen ermittelten Wahrscheinlichkeiten wesentlich besser die beobachteten Werte (Spalte 4) darstellen als die nach der Binomialverteilung berechneten. Insbesondere ist die Gesamtzahl der falsch positiven Wertesätze auch theoretisch kleiner als nach der Binomialverteilung zu erwarten.

Obwohl die Anzahl der falsch positiven Datensätze nach unseren Berechnungen geringer ist als nach der Binomialverteilung anzunehmen, erscheint dieser Anteil doch noch recht hoch. Das liegt zweifellos an der univariaten Betrachtungsweise, die den wirklichen Verhältnissen nicht gerecht werden kann. Es zeigt sich damit die Notwendigkeit der Definition multivariater Referenzbereiche. Dieses geschieht durch die Einführung von Testgrößen, die mehrere Variable gleichzeitig enthalten, wie kürzlich an einem Beispiel dargelegt wurde (11).

Faßt man etwa p standardisierte Variable $y_1, \ldots, y_p$ zu einem Vektor y zusammen, so empfiehlt sich als Testgröße nach *Hotelling* (7):

$$T^2 = y' C^{-1} y. \qquad \text{(Gl. 10)}$$

Dabei ist C die Korrelationsmatrix und y' der transponierte Datenvektor y. Dann ergibt sich als kritischer Wert für T^2:

$$T_0^2 = \frac{(N-1)\,p}{N-p} F_{p,N-p}(1-\alpha). \qquad \text{(Gl. 11)}$$

N ist dabei die Anzahl der Referenzpersonen, F kann einer Tabelle für die F-Verteilung entnommen werden. Je größer also das beobachtete T^2 ist, um so unplausibler bzw. pathologischer ist der gegebene Wertsatz (7). Hieraus ergeben sich verschiedene Vorteile:

1. Es fallen jetzt auch pathologische Wertekonstellationen auf, die bei univariater Betrachtungsweise unentdeckt geblieben wären. Abbildung 2 zeigt als Beispiel den Zusammenhang von Calcium- und Albuminwerten mit einer Korrelation von r = 0,61. Ein Patient, dessen Wertepaar an der durch einen Stern gekennzeichneten Stelle liegt, würde in den einzelnen Werten nicht auffallen. Trotzdem ist die beobachtete Konstellation von Werten unplausibel oder pathologisch.

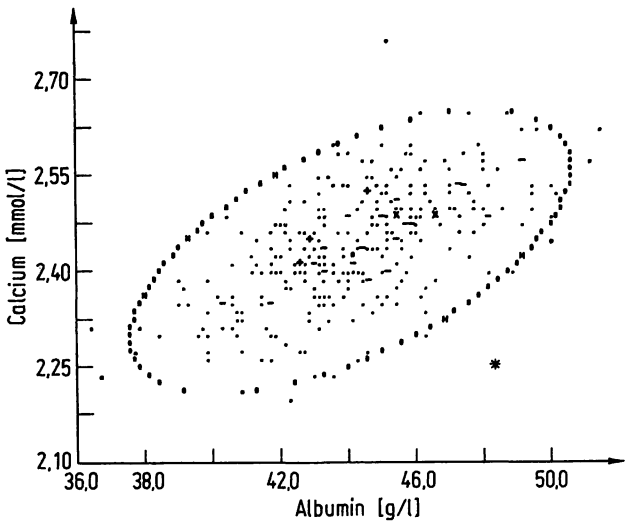

Abb. 2. Korrelation der Konzentrationen von Albumin (g/l) und Calcium (mmol/l) im Serum von Referenzpersonen. Eingezeichnet ist die Streuungsellipse ($\alpha = 0{,}95$). Der Korrelationskoeffizient (*Pearson*) beträgt r = 0,61.

2. Es wird eine beträchtliche Reduktion der falsch positiven Wertesätze erreicht. Definiert man nämlich einen falsch positiven Wertesatz dadurch, daß die Testgröße T^2 einen kritischen Wert $T_0^2(1-\alpha)$ überschreitet, so ist die Wahrscheinlichkeit für das Auftreten eines solchen Wertesatzes gerade gleich $P = \alpha$. Wählt man z. B. $\alpha = 0{,}05$, so ist die Anzahl der falsch positiven Wertesätze gleich 5% und zwar unabhängig von der Zahl der untersuchten Kenngrößen.

Wir haben die T^2-Werte unserer Referenzpersonen nach Gl. 10 berechnet. Abbildung 3 zeigt die Verteilung der Werte im Vergleich zur F-Verteilung. Obwohl die Verteilungen einiger Kenngrößen wie z. B. Bilirubin und Kreatinin deutlich rechtsschief sind, ergibt sich doch eine recht gute Übereinstimmung. Auch aus diesem Grund halten wir die beschriebene Testmethode für gut praktikabel. Bei unseren Referenzpersonen reduzierte sich z. B. die Anzahl der falsch positiven Profile auf 6,7%.

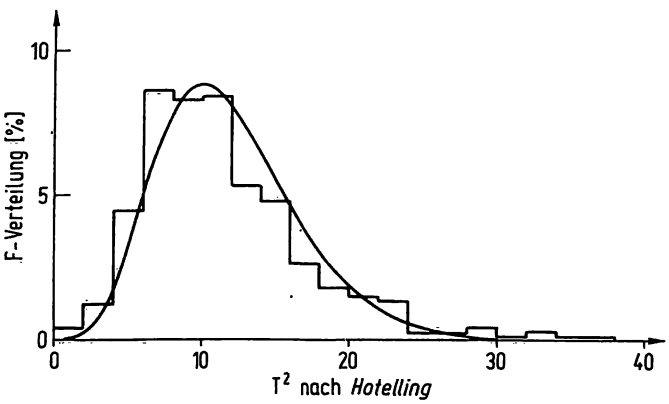

Abb. 3. Verteilung der Werte von *Hotellings* T^2 bei Referenzpersonen im Vergleich mit der F-Verteilung.

Für die praktische Berechnung der multivariaten Testgröße kann folgendermaßen vorgegangen werden:

Vor der Einzelfallanalyse muß man von den Werten eines geeigneten Referenzkollektivs ausgehen, die für das weitere Vorgehen zu Grunde gelegt werden. Für das weitere Vorgehen bieten sich zwei Möglichkeiten an, die mathematisch äquivalent sind.

1. Man berechnet den Mittelwertsvektor $\bar{x}$ und die Kovarianzmatrix V des Referenzkollektivs. Will man einen Patientenvektor x_p prüfen, so ergibt sich

$$T^2 = (x_p - \bar{x})' V^{-1} (x_p - \bar{x}) \qquad \text{(Gl. 12)}$$

als quadratische Form.

2. Man geht von der Korrelationsmatrix C der Referenzpersonen aus, muß dann aber den Patientenvektor standardisieren. Für die Komponenten y_i des standardisierten Datenvektors gilt also:

$$y_i = \frac{x_i - \bar{x}_i}{s_i} \qquad \text{(Gl. 13)}$$

T^2 ergibt sich dann nach Gl. 10.

Es sei angemerkt, daß die Berechnung der Kovarianz- bzw. Korrelationsmatrix und ihrer Inversen nur einmal durchgeführt werden muß und daß danach die Bestimmung der Testgröße keinerlei Schwierigkeiten macht und auf jedem programmierbaren Kleinrechner bzw. direkt on-line bei Verfügbarkeit eines Labor-Datenverarbeitungssystems vorgenommen werden kann.

Einschränkend und grundsätzlich ist zur Methodik noch folgendes zu sagen: Die theoretische Herleitung gilt nur für den Fall, daß die betrachteten Größen einer multivariaten Normalverteilung entstammen. In Wirklichkeit sind jedoch alle Größen mehr oder weniger nicht-normal verteilt. Diese Tatsache wirkt sich um so gravierender aus, je höher die Dimension des betrachteten Kenngrößenraumes ist. Wir schlagen daher vor, mit mehreren Testgrößen zu arbeiten, die jeweils etwa drei bis vier Kenngrößen enthalten. Dadurch wird zwar einerseits die Zahl der falsch positiven Werte wieder erhöht, andererseits besitzen die einzelnen Testgrößen bei geeigneter Kenngrößenauswahl eine größere Spezifität als eine globale Testgröße. Obwohl bei niedrigeren Dimensionen Abweichungen von der Normalität der Verteilungen sich nicht so gravierend auswirken, könnte man doch bei stark schiefen Verteilungen daran denken, die Werte vor der Rechnung zu transformieren (z. B. den Logarithmus zu nehmen), um so eine bessere Annäherung an eine Normalverteilung zu erreichen.

Die Zusammenstellung verschiedener ausgesuchter Kenngrößen zu einer Testgröße kann jedoch nicht nur nach statistischen Gesichtspunkten erfolgen, sondern muß auch nach medizinischen Kriterien vorgenommen werden. Es empfiehlt sich, darauf zu achten, daß in dem betrachteten Wertesatz einige signifikante Korrelationen vorhanden sind. Die den Daten innewohnende Redundanz, die sich in den Korrelationen zwischen den Variablen zeigt, läßt sich auf diese Weise auch für Plausibilitätsbetrachtungen nutzbar machen.

Literatur

1. Harm, K., Rehpenning, W., Domesle, A. & Voigt, K. D. (1979), diese Z. *17*, 517–522.
2. Büttner, J. (1977), diese Z. *15*, 1–12.
3. Morgenstern, D. (1964), Einführung in die Wahrscheinlichkeitsrechnung und mathematische Statistik, 8–9, Springer, Berlin–Göttingen–Heidelberg.
4. Buret, J., Monfort, F. & Marthoz, J. (1976), in: Organisation des laboratoires et interprétation des résultats – Biologie prospective. Comptes rendus du troisième colloque international, Pont-à-Mousson 1975 (Hrsg. Siest, G.), S. 751–754, L'Expansion Scientifique Française, Paris.
5. Grams, R. R., Johnson, E. A. & Benson, E. S. (1972), Amer. J. Clin. Pathol. *58*, 188–200.
6. Harm, K. (1974), Med. Progr. Technol. *3*, 45–55.
7. Anderson, T. W. (1958), An introduction to multivariate statistical analysis, S. 107–108, John Wiley & Sons, New York–London–Sydney.
8. Feller, W. (1957), An introduction to probability theory and its applications, S. 88–91, John Wiley & Sons, New York–London.
9. Piessens, R. (1973), Angewandte Informatik, 399–401.
10. Ralston, A. & Wilf, H. S. (1969), Mathematische Methoden für Digitalrechner II, S. 230–247, Oldenbourg, München–Wien.
11. Bernhardt, W., Weisner, B. & Rehpenning, W. (1978), diese Z. *16*, 435–439.

Dr. rer. nat. W. Rehpenning
Universitäts-Krankenhaus Eppendorf
Abteilung für Mathematik in der Medizin
Martinistraße 52
D-2000 Hamburg 20